Erschienen in: Kämper, Heidrun/Eichinger, Ludwig M. (Hrsg.): Sprach-Perspektiven. Germanistische Linguistik und das Institut für Deutsche Sprache. - Tübingen: Narr, 2007. S. 493-509. (Studien zur Deutschen Sprache 40)

Kathrin Steyer/Meike Lauer

„Corpus-Driven": Linguistische Interpretation von Kookkurrenzbeziehungen

Der Beitrag zeigt, auf welch grundlegende Weise das Paradigma der *Corpus-Driven-Linguistics* (CDL)[1] die linguistische Beschreibung sprachlichen Usus auf der Basis mathematisch-statistischer Clusteringverfahren bestimmt. Es soll deutlich werden, wie sich diese Prämissen im Forschungsschwerpunkt zur linguistischen Systematisierung und Interpretation von Kookkurrenzdaten manifestieren.[2]

Korpusgesteuert vorzugehen bedeutet, sich nach folgenden empirischen Grundprinzipien zu richten:

- Beobachtung der Sprachdaten;
- Akzeptanz aller Evidenzen;
- Bildung von Hypothesen auf Basis der Evidenzen;
- empirisches Prüfen der Hypothesen (Zusammenspiel von Induktion und Deduktion);
- deskriptive Aussagen, die die Evidenzen reflektieren;
- Generalisation im Sinne von Gebrauchsregeln.

Die linguistische Interpretation der Kookkurrenzdaten[3] erfolgt also konsequent nachgelagert (a posteriori). A posteriori heißt jedoch nicht vorausset-

---

1 Vgl. u.a. Sinclair (1991), Tognini-Bonelli (2001).

2 Wesentliche Anregungen und Impulse verdanken wir der Zusammenarbeit und vielen Diskussionen mit Cyril Belica. Er ist nicht nur für den Ausbau der IDS-Korpora der geschriebenen Sprache in ihrer heutigen Größe und Qualität verantwortlich, sondern vor allem auch für die elaborierte Erforschung und Entwicklung von korpusorientierten Erschließungsmethoden (vgl. u.a. Belica 1995, 2001-2006; Perkuhn et al. 2006).

3 An dieser Stelle kann nicht auf die Forschungsliteratur zu Kollokationen/Wortverbindungen und Korpora etc. eingegangen werden. Einen sehr umfangreichen Überblick gibt Hausmann (2004). Zur Funktionsweise der Kookkurrenzanalyse `http://www.ids-mannheim.de/kt/misc/tutorial.html`; zur linguistischen Perspektive bei der Analyse von Kookkurrenzdaten vgl. Steyer (1998, 2002, 2003, 2004), Belica/Steyer (i.Dr.).

zungslos und theoriefrei. Ohne Zweifel haben wir Ordnungssysteme im Kopf, die je nach Datenbefund die Interpretation der Daten mit bedingen. Diese Ordnungssysteme sind nicht immer gebunden an linguistische Kategoriensysteme. Spezifische linguistische Kategorien bilden demzufolge nicht den erfahrungsunabhängig vorgelagerten Erklärungsrahmen für das zu Analysierende und Bewertende. Sie dienen höchstens als ein Mittel der 'a posteriori-Kommentierung', und zwar in den Fällen, in denen sich ein interpretiertes Phänomen mit vorhandenen linguistischen Kategorien fassen lassen kann. Prinzipiell ist es aus korpusgesteuerter Sicht sinnvoller, die beobachteten Phänomene zunächst theorienneutral heuristisch zu beschreiben, natürlich mit dem Anliegen, einerseits vorhandene Theorien und Modelle auszudifferenzieren und gegebenenfalls zu revidieren, andererseits aber auch mit dem Ziel, zu neuen theoretischen Einsichten über das System Sprache zu gelangen.

Bezogen auf die korpusgesteuerte Interpretation von Kookkurrenzclustern und -profilen kann bei Vorliegen eines empirischen Befundes keine Bewertung danach erfolgen, ob die berechneten kohäsiven Zusammenhänge überhaupt der eigenen (linguistischen) Erwartungshaltung entsprechen oder auch nicht, ob sie also 'gut' oder 'schlecht', 'transparent' oder 'idiomatisch' sind usw. (vgl. dazu Belica/Steyer (i.Dr.), Perkuhn/Belica 2006). Vor dem Hintergrund korpusgesteuerter Empirie verbieten sich folgerichtig Interpretationen wie z.B. eine solche: Ein Kookkurrenzcluster ist hervorgetreten, **weil** es eine Kollokation, eine Paarformel oder ein Idiom ist. Vielmehr hat der Rechner[4] statistisch auffällige sprachliche **Verwendungszusammenhänge** erkannt, völlig unabhängig davon, wie man sie im Nachhinein benennt und einordnet. Er hat sie in Cluster zusammengefasst. Dies sind sozusagen die einzigen 'objektiven' (Sprach-)Fakten. Die linguistische Bewertung dieser Sprachfakten ist immer subjektiv. Sie stellt die Perspektive des Interpretierenden zu den berechneten sprachlichen Phänomenen dar. Alle berechneten Cluster werden also gleichberechtigt betrachtet. Das Hauptziel ist, zu erklären, **worin** die Ähnlichkeit der Verwendungskontexte besteht, die der Rech-

[4] Wir verwenden das Bild eines agierenden, denkenden und handelnden Rechners nur zur Veranschaulichung. Natürlich führt der Rechner nur die Prozesse aus, die der Mensch implementiert hat.

ner in einem Cluster zusammengefasst hat. Wir fragen uns also: Was unterscheidet diese Verwendungsmuster von anderen? Worin besteht ihre distinktive Qualität? In einem zweiten Schritt interessieren wir uns für Vertreter ähnlicher Art, also dafür, ob andere Cluster verwandte Eigenschaften aufweisen und so in Clustergruppen zusammengefasst werden. Schließlich geht es uns um die Frage, welche ab-strakteren Muster diesen einzelnen Gruppen von zusammengefassten Clustern zu Grunde liegen und wie diese im System Sprache miteinander vernetzt sind.

Anhand der Interpretation des Kookkurrenzprofils des Adjektivs *gesund* seien die eben beschriebenen Prämissen illustriert: Viele syntagmatische Muster des Kookkurrenzprofils lassen sich im herkömmlichen Sinne als feste, lexikalisierte Wortverbindungen interpretieren, z.B. *gesund und munter, auf gesunden Beinen stehen, in einem gesunden Körper wohnt ein gesunder Geist* oder *Lachen ist gesund.* Dafür könnten Kategorien der Phraseologie wie [ZWILLINGSFORMEL], [ZITAT] oder [ROUTINEFORMEL/SLOGAN] als 'Kommentar' dienen. Der in der Kollokations- und Idiomatikforschung nach wie vor präferierte und allgemein akzeptierte Zugang zur Systematisierung von lexikalischen Partnern eines Bezugswortes ist eine Ordnung nach morphosyntaktischen Kriterien.[5] Im Kookkurrenzprofil des Adjektivs *gesund* finden sich naturgemäß eine Vielzahl nominaler Kookkurrenzpartner.

(1)

Nominale Partner (nach LLR-Wert)[6]

Menschenverstand, *Ernährung*, *Volksempfinden*, *Menschen*, *Menschenverstandes*, *Lebensweise*, *Körper*, *Mischung*, *Umwelt*, *Portion*, *Basis*, *Gewebe*, *Kind*, *Geist*, *Lebensmittel*, *Essen*, [...] *Hauptsache* usw.

Schon ein kurzer Blick auf dieses Wortklassenfeld macht deutlich, dass diese Nomina sehr verschiedene Verwendungszusammenhänge indizieren und eine solche Systematisierung allein nicht besonders aussagekräftig ist. Viele

---

[5] Verwiesen sei hier nur auf die frühen Arbeiten von Hausmann, z.B. (1984).

[6] LLR-Wert bedeutet *log-likelihood-ratio*. Weitere Informationen zur Funktionsweise im Tutorial zur Kookkurrenzanalyse unter `http://www.ids-mannheim.de/kt/misc/tutorial.html` (Stand: November 2006).

dieser nominalen Kookkurrenzpartner lassen sich relativ unkompliziert in semantischen Feldern (heuristischen Kookkurrenzfeldern) zusammenfassen, die jeweils einen typischen Verwendungszusammenhang abbilden.

(2)

Heuristische Kookkurrenzfelder zu *gesund* (nur nominale Partner)

(Links sind die primären Kookkurrenzpartner aufgeführt, rechts die errechneten syntagmatischen Muster in der Darstellungsform der Kookkurrenzanalyse.)

[KONNOTATIVE QUALIFIZIERUNG]

Mischung	*eine gesunde* [...] *Mischung aus...*
Ehrgeiz	*gesunden Mischung aus ... Ehrgeiz*
Portion	*eine gesunde* [...] *Portion*
Ehrgeiz	*eine|Eine gesunde Portion Ehrgeiz*
Humor	*mit|und einer gesunden Portion Humor*
Skepsis	*eine gesunde Portion* [...] *Skepsis gegenüber ...*

[LEBENSWEISE]

Essen	*gesundes* [...] *Essen*
Bewegung	*Bewegung* [*und*] *gesundes Essen* [*auch gut*] *schmecken kann*
schmecken	*dass gesundes Essen* [*auch gut*] *schmecken kann*
Lebensmittel	gesunde [...] Lebensmittel
Tips [sic!]	*Tips und für|Ratschläge ein gesundes Leben*
Alter	*Alter gesund und körperlich fit*

[ÖKONOMISCH]

Füße	*auf wirtschaftlich gesunde Füße gestellt|zu stellen*
Füße	*Verein auf wirtschaftlich gesunde Füße*
Unternehmen	*ein wirtschaftlich* [...] *gesundes* [...] *Unternehmen zu ...*

[VERNUNFT]

Menschenverstand	*den gesunden* [...] *Menschenverstand*
Volksempfinden	*das gesunde* [...] *Volksempfinden*
Rechtsempfinden	*mit dem|einem gesunden Rechtsempfinden des ...*
Hausverstand	*den gesunden Hausverstand*

[KEINE GRUPPE]

Hauptsache	*Hauptsache* [... *ist*] *gesund*

Die Grenzen morphosyntaktisch motivierter Systematiken in Hinblick auf die Erhellung des typischen Gebrauchs werden z.B. bei jenen Nomina deutlich, die auf den ersten Blick nicht diese Funktion der semantischen Ausdifferenzierung erfüllen oder nicht mit einer linguistischen Kategorie z.B. [IDIOMATISCH] zu attribuieren sind, wie das beim in der letzten Gruppe zum nominalen Kookkurrenzpartner *Hauptsache* der Fall ist. Anhand der – ebenfalls mathematisch gewonnenen – syntagmatischen Mustersets

(3)

Hauptsache ich bin [...] *gesund*

Hauptsache [*man ist*] *gesund sagt*

Hauptsache [... *ist*] *gesund*

können wir erkennen, dass es sich um die Formel *Hauptsache gesund!* handelt. Die Einordnung in ein wortklassenbezogenes Kookkurrenzfeld nebengeordnet zu solchen Nomina wie *Lebensmittel*, *Füße*, *Ehrgeiz* und *Menschenverstand* bildet also hier in keiner Weise die Gebrauchsbesonderheit dieses Kookkurrenzclusters ab. Auf der Suche nach den zu Grunde liegenden Mustern analysieren wir das Kookkurrenzprofil von *Hauptsache,* indem wir eine so genannte Reziprokanalyse, nämlich die Kookkurrenzanalyse des Kookkurrenzpartners, durchführen. Die folgende Abbildung zeigt einen Ausschnitt aus dem Kookkurrenzprofil[7] von *Hauptsache*:

---

[7] Zur Aussagekraft der Platzhalter, Klammern usw. vgl. Tutorial zur Kookkurrenzanalyse unter `http://www.ids-mannheim.de/kt/misc/tutorial.html` (Stand: November 2006).

(4)

54% ist egal Hauptsache [wir] gewinnen
75% egal [...] wohin [...] Hauptsache
100% ist mir völlig egal [...] Hauptsache
93% Katze schwarz [oder weiß] ist Hauptsache sie fängt Mäuse
100% Hauptsache die Kasse stimmt
100% Hauptsache die Kohle stimmt
100% Hauptsache die Quote stimmt
100% Egal wohin Hauptsache weg
85% nach dem Motto [...] Hauptsache
66% Hauptsache billig ... die Devise
90% Hauptsache [...] billig
100% Hauptsache der Rubel rollt
100% Na ja Hauptsache Ihr könnt skifahren
88% nicht so wichtig [...] Hauptsache die

Analysiert man nun die syntagmatischen Einbettungen im Vor- und Nachfeld des Bezugswortes *Hauptsache,* lässt sich folgendes Grundmuster bestimmen:

(5)

[FESTSTELLUNG$_{changierend}$]	– *Hauptsache* –	[FESTSTELLUNG$_{positiv}$]
Katze [*schwarz oder weiß ist*]	***Hauptsache***	*sie fängt Mäuse*
Egal ob Mailand oder Madrid	***Hauptsache***	*Italien* (sic!)
Ganz egal [...] *wohin*	***Hauptsache***	*weg*
gewaltfrei oder militant	***Hauptsache***	*Widerstand*

Beispiele für vorangestellte Feststellung:

- *ist wurscht*
- *ist egal*
- *Aber was/Was soll's*

Beispiele für nachgestellte Feststellung:

- *die Kasse stimmt*
- *gesund*
- *es macht Spaß*

Die pragmatische Funktion dieser syntagmatischen Einbettungsmuster von *Hauptsache* ist die Zurückweisung einer wirklichen oder antizipierten negativen Bewertung. Der Sprecher äußert zunächst diese negative Bewertung oder formuliert Sachverhalte indifferent-changierend, um diese Aussage dann positiv aufzulösen. *Hauptsache* fungiert hier als 'argumentativer' Konnektor: Ein Argument wird stärker als ein anderes gewichtet. Rein kompetenzbasiert hätten wir dem Nomen *Hauptsache* wahrscheinlich keine Konnektorenfunktion zugeschrieben, dies konnte nur '*corpus-driven*' erfolgen. Interessant wäre nun zu fragen, ob es weitere Kandidaten gibt, die diese pragmatische Funktion erfüllen, obwohl sie völlig anders versprachlicht sind als die typischerweise als Konnektoren angesetzten lexikalischen Einheiten wie *und, oder, d.h., dass, ob, als ob, falls, da, weil, wobei, angenommen, außer, es sei denn, als, wie ...*[8] Die Hypothese, dass es noch viele andere sprachliche Formen gibt, die solche Konnektorenfunktionen erfüllen und die von der bisherigen Grammatikforschung nicht berücksichtigt wurden, lässt sich auf lange Sicht nur korpusgesteuert stützen und belegen. Man muss weitere Befunde sammeln. Andere sprachliche Formen mit dieser Funktion kann man nur bei systematischer Analyse entdecken, rein kompetenzbasiert sind sie kaum vorauszusagen.

Ein zweites Beispiel aus dem Kookkurrenzprofil des Adjektivs *gesund* sei angeführt: Ein signifikanter Kookkurrenzpartner des Adjektivs *gesund* ist das Nomen *Menschenverstand*. Es lässt sich in ein Kookkurrenzfeld einordnen, das die Bedeutung von *gesund* im Sinne von 'vernünftig, ausgewogen' fokussiert. Andere zu diesem Feld gehörende Partner sind z.B.: *Rechtsempfinden, Volksempfinden, Hausverstand,*[9] *Realismus, Humor, Sachver-*

---

[8] Vgl. dazu den Abschnitt zu Konnektoren im grammatischen Informationssystem des IDS *grammis*: `http://hypermedia.ids-mannheim.de/pls/public/termwb.ansicht?v_app=g&v_id=123` (Stand: November 2006).

[9] *Hausverstand* ist nur in den österreichischen Korpora belegt und wird dort in einem ähnlichen Sinne wie *Menschenverstand* verwendet, wobei *Hausverstand* besonders für das

stand, Vernunft. Analysiert man dann wiederum die Kookkurrenzprofile dieser eben genannten Kookkurrenzpartner von *gesund*, fällt eine Häufung von syntagmatischen Mustern auf, die miteinander verwandt zu sein scheinen, ohne dass sie bisher in einer linguistischen Klasse[10] als zusammengehörig definiert wurden, und zwar mehrgliedrige Quantoren wie *eine [gesunde, gehörige ...] Portion, ein gewisses Maß an, eine/keine Spur/ein/kein Hauch von, eine Prise, ein Schuss, ein Anflug von, mit viel, gespickt mit.* Isoliert betrachtet kann man den beteiligten Nomina *Portion, Spur, Prise, Schuss* oder *Anflug* keine direkten 'Verwandtschaftsbeziehungen' zuschreiben.[11] Einzig dem Paar *Portion* und *Prise* ist eine gewisse Synonymie zu unterstellen. Erst durch die Verwendungszusammenhänge im Vergleich ihrer Kookkurrenzprofile sind sie als Vetreter einer eigenen Ausdrucksklasse zu interpretieren, die der **mehrgliedrigen konnotativen Quantoren**. Korpusgesteuert bedeutet in diesem Fall, dass wir erstens diese Beobachtung selbstverständlich festhalten und dass wir zweitens weiter verfolgen, ob wir auf andere Vertreter stoßen, die diese neue Ausdrucksklasse rechtfertigen. Wieder haben wir eigentlich ein semantisches Feld eines Bezugswortes betrachtet und sind dabei aber auf Muster gestoßen, die mit der lexikalischen Ausgangseinheit nichts mehr zu tun haben.

Beide Beispiele haben deutlich gemacht, dass die Analyse des Kookkurrenzprofils des Bezugswortes *gesund* nur der Ausgangspunkt für eine Vielzahl anderer Erkenntnisse war. Die lexikalische Einheit stellte quasi nur das „Eingangstor" für linguistische Beobachtungen auf allen Ebenen des Sprachsystems dar.

---

praktische, realitäts- und naturnahe Denken 'kleiner Leute' und ihr entsprechendes pragmatisches Urteilsvermögen gebräuchlich ist, oft auch im Gegensatz zu wissenschaftlichem Denken.

[10] Wir beziehen uns hier auf die Klassifikationen der inzwischen zu den Referenzwerken der Mehrwortforschung gewordenen Monografien von Fleischer (1997) und Burger (2003). Natürlich können wir nicht ausschließen, dass sich Einzelanalysen dem Phänomen möglicherweise schon gewidmet haben.

[11] Diese zunächst kompetenzbasiert angenommene 'Nicht-Verwandtschaft' konnten wir ebenfalls mit Hilfe eines in die Kookkurenzdatenbank CCDB integrierten automatischen Verfahrens, des von Cyril Belica entwickelten Moduls *„Similar Collocation Profiles"* (*automatische Ermittlung verwandter Kookkurrenzprofile*) verifizieren.

Die wichtigsten Analysefragen lassen sich wie folgt zusammenfassen:

- Welche Binnenstruktur weist das Kookkurrenzcluster auf?
- Welches Muster liegt diesem Syntagma zugrunde?
- Welche anderen Vertreter konstituieren dieses Grundmuster?
- Welche typischen Variationen und Gebrauchsrestriktionen lassen sich erkennen (u.a. Festigkeitsgrad)?
- In welche typischen Umgebungsmuster ist das Syntagma eingebettet?
- Gibt es andere Kookkurrenzcluster, zu denen dieses Cluster in Verwandtschaft steht?
- Welche pragmatische Funktion erfüllt das Gebrauchssyntagma?
- Lassen sich diese Interpretationen zu anderen Fällen ähnlicher Art zuordnen?

Der hier vorgestellte Forschungsschwerpunkt widmet sich bei der Beschreibung usueller Wortverbindungen des Deutschen in ihrem aktuellen Gebrauch und ihrer inneren Systematik also vorrangig diesen Fragen. Usuelle Wortverbindungen[12] sind als rekurrente Muster des Sprachgebrauchs zu interpretieren, die sich in Kookkurrenzprofilen vor allem im hochfrequenten Bereich manifestieren. Indem wir die Systematik dieser Sprachgebrauchsmuster aufdecken, entschlüsseln wir zentrale Konstitutionsformen von Sprache. Von besonderem Stellenwert ist das Konzept von Wortverbindungen als **Gebrauchssyntagmen**. Es geht uns also in diesem Forschungszusammenhang nicht um den Zugang zu Bedeutung und Gebrauch einzelner sprachlicher Ausdrücke über die affinen Kookkurrenzpartner im Sinne von Firth,[13] sondern um die Analyse von Gebrauchssyntagmen als holistische Einheiten[14] sprachlicher Kommunikation schlechthin. Diese Syntagmen sind durch massenhaften Gebrauch entstanden und werden als Bausteine im Sprach-

---

[12] Zum Konzept *Usuelle Wortverbindungen* vgl. Steyer (2000).

[13] Vgl. Firth (1968); zum Konzept der Anwendung der statistischen Kookkurrenzanalyse für die lexikografische Praxis des Deutschen (u.a. der heuristischen Kookkurrenzfelder) vgl. Steyer u.a. (2002, 2003, 2004), Belica/Steyer (i.Dr.).

[14] Auch Siepmann verwendet in diesem Zusammenhang den Terminus 'holistisch'. Kollokationen definiert er folgendermaßen: „a collocation is any **holistic** lexical, lexicogrammatical or semantic unit normally composed of two or more words which exhibits minimal recurrence within a particular discourse community" (Siepmann 2005, S. 438; Hervorhebung durch die Autorinnen).

gebrauch aktualisiert und eingesetzt.[15] Der Rechner erkennt sie in genau dieser Form als rekurrent, z.B.:

(6) [STRUKTURELL-BINÄR]
zunehmend schwieriger
groß angelegt
neu aufrollen

(7) [STRUKTURELL-TRINÄR]
das große Zittern
in höchster Alarmbereitschaft
es fehlt an
kein Grund zur Entwarnung

(8a) [FUNKTIONAL]
sich fragen lassen müssen
manchmal fragt man sich schon
immer mehr fragen sich
anders gefragt

Viele derartige syntagmatische Muster scheinen aus 'regelgeleiteter' Sicht einer Phrasen- und Satzperspektive trivial oder nur unvollständig, in manchen Fällen gar idiosynkratisch. Aber der Rechner hat sie genau in dieser Form als zusammenhängendes Verwendungscluster erkannt. Ihre Restriktion liegt im wiederkehrenden Gebrauch genau dieser sprachlichen Formen.[16] So liegt

---

[15] Stubbs betont, dass es sich bei solchen Kombinationen um „Normen des Sprachgebrauchs" handelt (Stubbs 1997, S. 157).

[16] Der entsprechende sprachliche Prozess findet sich schon in der Sprachtheorie von Feilke wieder. Diese Theorie, besonders sein Konzept der idiomatischen Prägung, stellt u.E. eine der wichtigsten theoretischen Begründungen korpuslinguistischer Empirie – und das quasi in Vorwegnahme der *Corpus-Driven Linguistics* – dar (Feilke 1996).
Wir sehen hier interessante Querverbindungen zum Konzept kommunikativer Minimaleinheiten, das in der IDS-Grammatik entwickelt wurde (Zifonun/Hoffmann/Strecker 1997, Bd. 1, S. 85ff.). Verwiesen sei auch auf 'funktionale Einheiten', die im IDS-Projekt 'Eigenschaften des gesprochenen Deutsch' untersucht und beschrieben wurden (Fiehler/Barden/Elstermann/Kraft 2004, S. 204ff.).

die Gebundenheit der Gebrauchssyntagmen in (7) *in höchster Alarmbereitschaft* in der im Korpus fast ausschließlich vorkommenden Superlativform, bei *es fehlt an* ist es das obligatorische Präpositionalkomplement *an*,[17] bei *kein Grund zur Entwarnung* das Negationsmuster. Die Gruppe (8b) verdeutlicht sehr anschaulich, dass die ermittelten Gebrauchssyntagmen in genau dieser grammatischen Form spezifische kommunikative Funktionen erfüllen:

(8b) *sich fragen lassen müssen* (Warnung/Kritik)
manchmal fragt man sich schon (Zweifel/Kritik)
immer mehr fragen sich (Zweifel/Akzeptanzstützung)
anders gefragt (metakommunikative Textstrukturierung)

Uns interessiert nun, ob es weitere sprachliche Realisierungen solcher kommunikativer Funktionen durch andere Gebrauchssyntagmen gibt, wie z.B. im Kookkurrenzprofil von *fragen:*

(9) [ZWEIFEL/KRITIK]
viele fragen sich warum ausgerechnet
sich fragen warum
sich ernsthaft fragen
oft habe ich mich gefragt
ungläubig fragen
sich allen Ernstes fragen
fragt sich bloß
fragen die Leute
mag man fragen
mag sich mancher gefragt haben
immer mehr fragen sich

[17] Auch in VALBU ist diese Konstruktion als eigenständiges Lemma gleichberechtigt zum Eintrag des Verbs *fehlen* angesetzt (Schumacher/Kubczak/Schmidt/de Ruiter 2004, S. 373ff.).

Unsere wichtigsten Analyseinteressen lassen sich wie folgt beschreiben:

- Clustergruppen zusammengefasst nach inhaltlichen Kriterien
  Beispiel (1)
- Abstrakte Muster (Platzhalter und konkrete Realisierungen)
  Beispiele (2)-(5)
- neue Vertreter für bekannte Ausdrucksklassen
  Beispiele (3)-(5)
- neue Ausdrucksklassen
  Beispiel: *gesund* → mehrgliedrige konnotative Quantoren
- Zusammenhänge zwischen grammatischer Form und kommunikativer Funktion
  Beispiele (6)-(9)

Perspektivisch sollen Systematisierungen von Kookkurrenzclustern in Form von hypertextuellen Wortverbindungsnetzen dargestellt werden (zum generischen Modell solcher Informationsnetze siehe den Beitrag von Rainer Perkuhn in diesem Band). Diese Netze bestehen aus Knoten (Informationseinheiten), die durch das jeweilige abstrakte Ausdrucksmodell bestimmt sind und in (teils hierarchischen) Netzstrukturen abgebildet werden. Im Mittelpunkt stehen dabei stets die **sprachlichen Daten** selbst. Sprachliche Daten bedeuten hier Cluster von Originaltextausschnitten, so wie sie die Kookkurrenzanalyse berechnet. Diese sollen in eine inhaltlich motivierte, erkenntnisleitende Ordnung gebracht werden, so dass man sich **primär** aus der „Ordnung der Dinge" selbst ein Bild vom typischen Gebrauch und den kontextuellen Gegebenheiten machen kann. Es soll somit also ein reflektierter und strukturierter Zugang zu sprachlichen Massendaten geschaffen werden. Zusätzlich werden Metadaten wie linguistische Erläuterungen oder Gebrauchskommentare hinzugefügt.

Um zu einer systematisierten Darstellung und Visualisierung in Netzen zu gelangen, ist eine **linguistische Annotationssystematik** unabdingbar – eine Annotationssystematik, die es erst ermöglicht, sprachliche Erscheinungen und Muster nach bestimmten Kriterien in Ausdrucksklassen und diese dann in Informationseinheiten (Knoten) zusammenzufassen. Mit der Konzipierung eines solchen Annotationsmodells scheint man vor einem fast unlösbaren Dilemma zu stehen: Einerseits soll ein solches Modell das grundlegende in-

duktive Prinzip der CDL nicht konterkarieren und also möglichst wenig sprachwissenschaftliches Wissen – sprich ganz bestimmte linguistische Kategorien und Modelle – als geltende Annotations'ontologie' vorgeben. Es soll das Erkennen und Darstellen neuer Klassen und abstrakterer Zusammenhänge höherer Ordnungen ermöglichen. Vor allem aber soll das Modell dynamisch sein und als heuristisches Instrument helfen, die sprachlichen Erscheinungen zu erkennen und darzustellen, die sich dem Blick des Linguisten bisher verschlossen haben und für die es folgerichtig noch keine Benennungen und Kategorien gibt. Andererseits sind für eine nachgelagerte Systematisierung von derartigen sprachlichen Massendaten und für eine hypertextuelle Präsentation wiederum konkrete Attribute/Labels unverzichtbar – Attribute, mit deren Hilfe die Cluster linguistisch 'kommentiert' werden können und die eine gewisse Formalisierbarkeit überhaupt erst möglich machen.[18]

Neben dem theoretischen Interesse an der Entschlüsselung sprachlicher Strukturen verfolgen wir auch konkrete anwendungsbezogene Ziele. So wird ein Schwerpunkt in der Aufbereitung solcher Erkenntnisse für die Bedürfnis-

[18] Annotationsmodelle in der automatischen Sprachverarbeitung sind dieser Problematik nicht in dem Maße ausgesetzt: Sprachliche Objekte/Daten werden ausgezeichnet nach einem vorgegebenen linguistischen Modell. Dieses Modell bildet dann auch die Basis für die 'Ordnung der Dinge' und die darauf basierenden Erkenntnisse in genau nur diesem linguistischen Paradigma, z.B. der zugrunde liegenden Grammatik. In diesem Zusammenhang sei auf zwei der wichtigsten Annotationsverfahren verwiesen: das Wortart-Tagging, das Tokens mit Hilfe eines so genannten Part-of-Speech-Tagger nach Wortarten klassifiziert, und das Verfahren des Parsings, das syntaktische Strukturen in der Hierarchie ihrer Konstituenten erfasst (vgl. Carstensen et al. 2004, S. 224ff.). All diesen Verfahren ist gemeinsam, dass sie linguistisches Wissen (syntaktische und semantische Modelle) a priori hinzufügen, damit die so ausgezeichneten Dokumente „mit informationstechnologischen Werkzeugen bearbeitet werden können [...]. Derart bearbeitete Dokumente stellen eine wertvolle Ressource sowohl für höhergeordnete Anwendungen wie z.B. dem Information Retrieval [...] oder der Lexikographie [...] als auch für grundlegende Analysen wie der Kollokationsextraktion oder dem Syntaxparsing [...] dar" (ebd., S. 218). Übergeordnetes Ziel ist hier aber das Erreichen von Ergebnissen, die sich möglichst in Regeln fassen lassen, und somit wiederum streng formalisierbar und für eine weitere automatische Verarbeitung geeignet sind.

se im Rahmen von 'Deutsch als Fremdsprache' liegen.[19] Den größten Nutzen sehen wir in der Möglichkeit der feinen Kontextspezifizierung, die durch solche Methoden wie die statistische Kookkurrenzanalyse in einer neuen Qualität möglich wird. Deshalb ist die Kookkurrenzanalyse aus unserer Sicht besonders für ein mittleres bis gehobenes Sprachniveau und besonders auch für alle Multiplikatoren des Deutschen wie Auslandsgermanisten, Lehrer, Übersetzer oder Lexikografen von besonderem empirischem Wert.

Es sollte deutlich geworden sein, dass das induktive Prinzip der CDL mit den am IDS entwickelten automatischen Verfahren sehr viel konsequenter und umfassender umgesetzt werden kann, als das der einzelne linguistische Forscher je leisten könnte, und dass ein iteratives Zusammenspiel zwischen automatischen und 'intellektuellen' Zugängen zu einer neuen Qualität in der Sprachbeschreibung führen kann. Unsere Forschungen sollen demzufolge langfristig auch die linguistische Theorienbildung unterstützen, vor allem durch eine Reflexion der Interaktion von automatischen sprachgebrauchsorientierten Methoden und der eher deduktiv orientierten Theoriediskussion in der 'Systemlinguistik'. Hier erhoffen wir uns, einen Impuls für eine unserer Meinung nach anstehende Selbstreflexion unseres Faches und seiner theoretischen und methodologischen Leitprinzipien geben zu können.

Wir stehen erst am Anfang.

---

[19] Das Projekt *Usuelle Wortverbindungen* unterhält und pflegt seit Jahren Kooperations- und Arbeitsbeziehungen mit auslandsgermanistischen Projekten, die in den kommenden Jahren intensiviert werden sollen, z.B. im Rahmen der Europäischen Gesellschaft für Phraseologie. Eine besonders enge und produktive Kooperation besteht mit dem Institut für Germanische Studien an der Karls-Universität Prag, dort vor allem mit dem „Deutsch-tschechischen akademischen Wörterbuch" (vgl. u.a. Steyer/Vachková i.Dr.).

Literatur

Belica, Cyril (1995): Statistische Kollokationsanalyse und Clustering. COSMAS-Korpusanalysemodul. Mannheim 1995. Internet: http://corpora.ids-mannheim.de (Stand: November 2006).

Belica, Cyril (2001-2006): Kookkurrenzdatenbank CCDB. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. Mannheim. http://corpora.ids-mannheim.de/ccdb/ (Stand: November 2006).

Belica, Cyril/Steyer, Kathrin (i.Dr.): Korpusanalytische Zugänge zu sprachlichem Usus. In: AUC (Acta Universitatis Carolinae), Germanistica Pragensia XX. Praha.

Burger, Harald (2003): Phraseologie. Eine Einführung am Beispiel des Deutschen. 2., überarb. Aufl. Berlin. (= Grundlagen der Germanistik 36).

Carstensen, Kai-Uwe/Ebert, Christian/Endriss, Cornelia/Jekat, Susanne/Klabunde, Ralf/Langer, Hagen (Hg.) (2004): Computerlinguistik und Sprachtechnologie. Eine Einführung. 2. überarb. u. erw. Aufl. München.

Feilke, Helmuth (1996): Sprache als soziale Gestalt. Ausdruck, Prägung und die Ordnung der sprachlichen Typik. Frankfurt a.M.

Fiehler, Reinhard/Barden, Birgit/Elstermann, Mechthild/Kraft, Barbara (2004): Eigenschaften gesprochener Sprache. Tübingen. (= Studien zur Deutschen Sprache 30).

Firth, John R. (1968): A Synopsis of Linguistic Theory 1930-1955. In: Palmer, Frank (Hg.): Selected Papers of J. R. Firth. Harlow. S. 168-205. [Zuerst ersch. in: Studies in Linguistic Analysis. Philological Society. Oxford: 1957.].

Fleischer, Wolfgang (1997): Phraseologie der deutschen Gegenwartssprache. 2., durchges. u. erg. Aufl. Tübingen.

Hanks, Patrick (2004): The Syntagmatics of Metaphor and Idiom. In: International Journal of Lexicography 17, 3, S. 245-274.

Hausmann, Franz Josef (1984): Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. In: Praxis des neusprachlichen Unterrichts 31, 4, S. 395-406.

Hausmann, Franz Josef (2004): Was sind eigentlich Kollokationen? In: Steyer (Hg.) 2004a, S. 309-334.

Perkuhn, Rainer/Belica, Cyril (2006): Korpuslinguistik – das unbekannte Wesen. In: Sprachreport 1/2006, S. 2-8.

Perkuhn, Rainer/Belica, Cyril/al-Wadi, Doris/Lauer, Meike/Steyer, Kathrin/Weiß, Christian (2006): Korpustechnologie am Institut für Deutsche Sprache. In: Schwitalla, Johannes/Wegstein, Werner (Hg.): Korpuslinguistik deutsch: synchron – diachron – kontrastiv. Würzburger Kolloquium 2003. 20.-23.3.2003. Universität Würzburg. Tübingen. S. 57-70.

Schumacher, Helmut/Kubczak, Jacqueline/Schmidt, Renate/de Ruiter, Vera (2004): VALBU – Valenzwörterbuch deutscher Verben. Tübingen. (= Studien zur deutschen Sprache 31).

Siepmann, Dirk (2005): Collocation, Colligation and Encoding Dictionaries. Part I: Lexicological Aspects. In: International Journal of Lexicography 18, 4, S. 409-443.

Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford.

Steyer, Kathrin (1998): Kollokationen als zentrales Übersetzungsproblem – Vorschläge für eine Kollokationsdatenbank Deutsch-Französisch/Französisch-Deutsch auf der Basis paralleler und vergleichbarer Korpora. In: Bresson, Daniel (Hg.): Lexikologie und Lexikographie Deutsch-Französisch. Aix-en-Provence. (= Cahiers d'Études Germaniques 35). S. 95-113.

Steyer, Kathrin (2000): Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten. In: Deutsche Sprache 2/2000, S. 101-125.

Steyer, Kathrin (2002): Wenn der Schwanz mit dem Hund wedelt. Zum linguistischen Erklärungspotenzial der korpusbasierten Kookkurrenzanalyse. In: Haß-Zumkehr, Ulrike/Kallmeyer, Werner/Zifonun, Gisela (Hg.): Ansichten der deutschen Sprache. Festschrift für Gerhard Stickel zum 65. Geburtstag. Tübingen. (= Studien zur Deutschen Sprache 25). S. 215-236.

Steyer, Kathrin (2003): Korpus, Statistik, Kookkurrenz. Lässt sich Idiomatisches „berechnen"? In: Burger, Harald/Häcki Buhofer, Annelies/Gréciano, Gertrud (Hg.): Flut von Texten – Vielfalt der Kulturen. Ascona 2001 zur Methodologie und Kulturspezifik der Phraseologie. Baltmannsweiler. (= Phraseologie und Parömiologie 14). S. 33-46.

Steyer, Kathrin (2004): Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: Steyer (Hg.), S. 87-116.

Steyer, Kathrin (Hg.) (2004): Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003. Berlin/New York.

Steyer, Kathrin/Vachková, Marie (i.Dr.): Kookkurrenzanalyse kontrastiv. Zum Nutzen von Korpusanalysemethoden für die bilinguale lexikografische Praxis am Beispiel des GDTAW. In: AUC (Acta Universitatis Carolinae), Germanistica Pragensia XX. Praha.

Stubbs, Michael (1997): „Eine Sprache idiomatisch sprechen“: Computer, Korpora, Kommunikative Kompetenz und Kultur. In: Mattheier, Klaus J. (Hg.): Norm und Variation. Frankfurt a.M. (= Forum Angewandte Linguistik 32). S. 151-167.

Tognini-Bonelli, Elena (2001): Corpus Linguistics at Work. Amsterdam/Philadelphia. (= Studies in Corpus Linguistics 6).

Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno (1997): Grammatik der deutschen Sprache. 3 Bde. Berlin/New York. (= Schriften des Instituts für deutsche Sprache 7.1-3).